Using Haplotype Mapping to Uncover the Missing Heritability: A Simulation Study

M. Shirali,[1] R. Pong-Wong,[2] S. Knott,[3] C. Hayward,[1] V. Vitart,[1] I. Rudan,[4, 5] H. Campbell,[5] N. Hastie,[1] A. Wright,[1] P. Navarro,[1] C. Haley[1, 2]

[1]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK, [2]The Roslin Institute and R (D) SVS, University of Edinburgh, Midlothian, UK, [3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK, [4]Croatian Centre for Global Health, Faculty of Medicine, University of Split, Split, Croatia, [5]Centre for Population Health Sciences, University of Edinburgh, Edinburgh, UK.

ABSTRACT: An improved regional mapping method is proposed based on haplotype information with haplotype blocks being used as analysis regions. Genomic data were simulated using *circa* 300K SNPs. The simulated phenotypes' heritability was 0.30 from which 0.05 was regional heritability. Twenty different regions of genome were selected to be trait-associated in the simulation. Four scenarios were used to generate regional variance with either one SNP, all SNPs, one haplotype or all haplotypes from the region being the causal variants. Regional genomic relationship matrices constructed with SNP-based or haplotype-based methods were used in a REML framework to estimate the variance explained by the region. Our results show that the proposed haplotype-based method always captures the effect of the regions, albeit with decreased performance with increasing block size for SNP-based variants. SNP-based methods often do not detect the effect of causal haplotype(s).

Key words: missing heritability; regional mapping method; haplotype blocks.

INTRODUCTION

Despite the success of genome-wide association studies in detecting common variants for quantitative traits, a huge portion of the genetic variance remains to be explained for most traits. The missing heritability may result from imperfect linkage disequilibrium (LD) between the genotyped SNPs and causal variants (Yang et al. (2010)). This low LD is likely to be largely due to causal variants having lower minor allele frequencies than genotyped SNPs. Therefore, an analytical approach is needed to capture causal variants with low minor allele frequencies.

In this study, a Haplotype Heritability Mapping method (HHM), which has been developed from the regional heritability mapping method of Nagamine et al. (2012), is presented as a novel approach to the analysis of genome-wide association data. The HHM method uses haplotype blocks in a region as the unit to estimate relationships and perform analysis and thus utilizes the LD information in the region. The HHM method exploits a mixed model framework with a genome-wide and a regional genomic relationship matrix in the model, to estimate trait variance associated with the region, potentially capturing the joint effects of both common and rare variants in that region.

MATERIALS AND METHODS

The simulation study was based on real genotype data obtained from a human population.

Population and SNP array information. Samples were available from three Southern European cohorts: from the city of Split and islands of Vis and Korcula on the Dalmatian coast of Croatia. The study was approved by the Ethical Committee of the Medical School, University of Zagreb and the Multi-Centre Research Ethics Committee for Scotland. All participants gave written informed consent. The samples were genotyped using 300K SNP genotyping arrays (Illumina Human Hap300 for Vis and Illumina CNV370 for Korcula and Split). Quality control procedures were performed per SNP and per individual. SNPs with minor allele frequency < 0.01, out of Hardy-Weinberg equilibrium ($P < 10^{-8}$) and with a call rate < 0.95 were discarded. Individuals with a call rate of < 0.95 were excluded. After quality control, 2186 individuals remained, and 267,136 autosomal SNPs, genotyped in all the populations, were used in our analysis.

Simulation data. Haplotypes were inferred from the available genotypic data using BEAGLE version 3.3.2 (Browning and Browning (2007)). These were used as base population haplotypes. For each replicate of the simulation, the 2186 base-population individuals were randomly selected as parents of generation 1. Subsequently random mating was simulated for the next 20 generations. Population size was kept constant over generations at the base population size. Genotypes from generation 20 were used in the current study. Ten replicates were performed.

We simulated phenotypes for which the total variance was 1 and the heritability was 0.30, of which 0.25 was polygenic variance for which all SNPs were assumed to have a very small effect on the phenotype and the remaining 0.05 was regional variance. To simulate the regional heritability, 20 different regions were selected. Ten of them contained the top 10 SNPs reported in a meta-analysis of HDL (Teslovich et al. (2010)) and the other 10 were control regions, each selected within the same chromosome as one of the reported hits.

The regions were haplotype blocks delimited by recombination hotspots with at least 5 centiMorgans per megabase (cM/Mb) based on the Genome Reference Consortium Human Build 37. The number of SNPs within the chosen blocks for this study was between 1 and 72. In

each region, four different alternative scenarios to simulate regional heritability were assumed. The regional effect was simulated using either one SNP (1S) in the region, all SNPs (AS) in the region (each one contributing equal variance), one randomly selected haplotype (1H) in the block or, finally, all haplotypes (AH) in the block. The effect of each haplotype on traits in the AH scenario was randomly selected from a normal distribution. In total 80 quantitative traits (20 regions and 4 scenarios per region) were simulated and 10 replicates performed.

Regional Haplotype Mapping (HHM) Method. To map trait-associated regions, HHM was performed based on the variance component method described by Nagamine et al. (2012). In the current study, instead of using fixed-size windows containing a constant number of SNPs as Nagamine et al. (2012), haplotype blocks were used as the regions for analysis. To determine the haplotype blocks we used block boundaries based on recombination hotspots with at least 5 cM/Mb recombination rate based on the Genome Reference Consortium Human Build 37. To calculate the regional genomic relationship matrices (RGRM), 4 different methods were used, two of which were based on SNP data and two based on haplotypes.

In the first SNP-based RGRM estimation method (SB1), the following equation was used for diagonal and off-diagonal elements of the RGRM.

$$IBS_{ij} = \frac{1}{S}\sum_{k=1}^{S}\frac{(O_{ik}-2P_k)(O_{jk}-2P_k)}{2P_k(1-P_k)}$$

where IBS_{ij} is the relationship between individual i and j, S is number of SNPs in the block, O_{ik} and O_{jk} are the genotypes of the i-th and j-th individuals at the k-th SNP (coded as 0, 1, and 2 for AA, AB and BB, respectively) and P_k is the frequency of k-th SNP (based on Aulchenko et al. (2007)).

In the second SNP-based RGRM estimation method (SB2) following equation was used:

$$IBS_{ij} = \frac{\sum_{K=1}^{S}(O_{ik}-2P_k)(O_{jk}-2P_k)}{\sum_{K=1}^{S}2P_k(1-P_k)}$$

(VanRaden (2007)).

In the first haplotype-based estimation RGRM method (HB1) we propose following equation:

$$IBS_{ij} = \frac{1}{H}\sum_{k=1}^{H}\frac{(O_{ik}-2P_k)(O_{jk}-2P_k)}{2P_k(1-P_k)}$$

where IBS_{ij} is the relationship between individual i and j, O_{ik} and O_{jk} are the genotypes of the i-th and j-th individuals at the k-th haplotype (coded as 0, 1, and 2 for H_hH_h, H_kH_h and H_kH_k, respectively, for k not equal to h) and P_k is frequency of k-th haplotype.

In the second haplotype-based RGRM estimation method (HB2) we propose:

$$IBS_{ij} = \frac{\sum_{K=1}^{S}(O_{ik}-2P_k)(O_{jk}-2P_k)}{\sum_{K=1}^{S}2P_k(1-P_k)}$$

where IBS_{ij} is the relationship between individual i and j, O_{ik} and O_{jk} are the genotypes of the i-th and j-th individuals at the k-th haplotype (coded as 0, 1, and 2 for H_hH_h, H_kH_h and H_kH_k, respectively, for k not equal to h) and P_k is frequency of k-th haplotype. When the block size is equal to 1, the SB1 and HB1 RGRMs are identical, as are the RGRMs for SB2 and HB2.

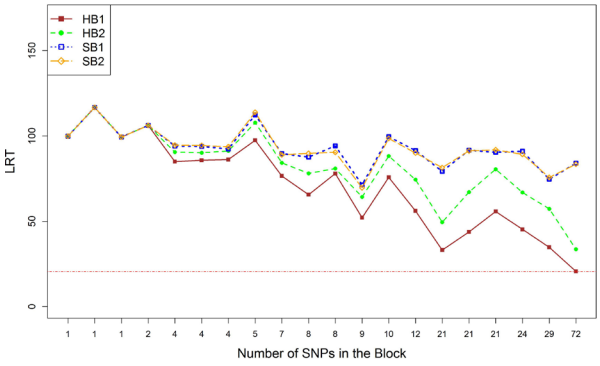

Figure 1. Average LRT for the 20 block regions ordered by size in 1S simulations.

After generating IBS matrices using custom-made scripts, REACTA version 0.9.7 (Cebamanos et al. (2014)) was used to solve the equations.

RESULTS AND DISCUSSION

1S and AS phenotype. 1S represents a situation in which there is a single SNP causal variant influencing the trait in the region of interest, whereas, AS has multiple causal variants each tagged by a SNP. The average likelihood ratio test (LRT) over all replicates for the 4 alternative analysis methods is presented in Figure 1 and Figure 2 for 1S (one SNP) and AS (all SNPs) phenotype simulations, respectively, for all 20 regions. In both cases the results suggest that SNP-based RGRM methods perform better than haplotype-based RGRM methods and that they have relatively constant performance to detect the simulated effect with limited decline in the LRT as block size increases. The haplotype-based RGRM methods can capture the QTL effects in all 20 regions, but their performance decreases with increasing block sizes. When the regional genetic variance is explained by SNP(s) in long haplotype blocks, unrelated individuals containing the same SNP(s) but having different haplotypes are considered as unrelated in haplotype-based RGRM methods; however, they are considered related using SNP-based RGRM methods, which could explain the difference in power between methods.

1H and AH phenotype. 1H and AH represent situations where the causal variants are not in strong LD

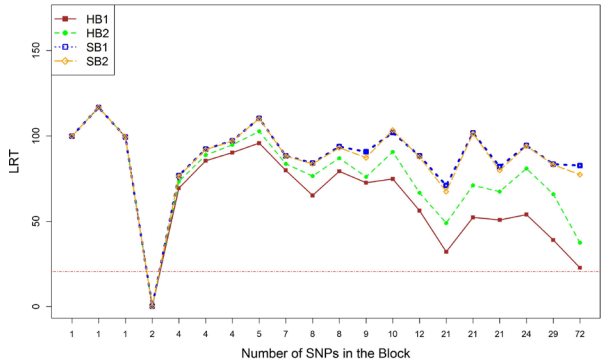

Figure 2. Average LRT for the 20 block regions ordered by size in AS simulations.

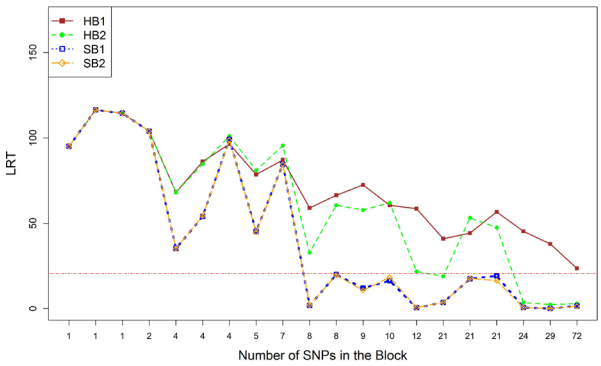

Figure 3. Average LRT for the 20 block regions ordered by size in 1H simulations.

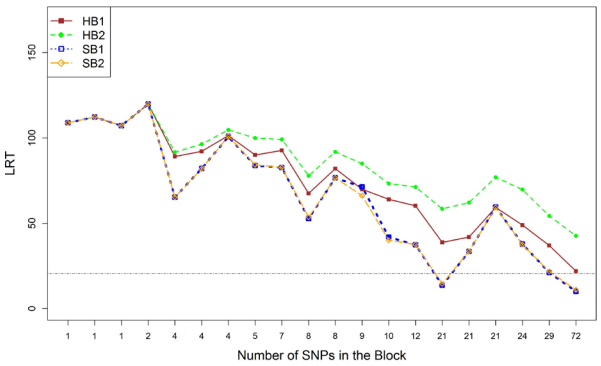

Figure 4. Average LRT for the 20 block regions ordered by size in AH simulations.

with any single SNP. For 1H there is a single causal variant whereas for AH there are multiple causal variants such that each haplotype in the region is associated with a different effect. The average likelihood ratio test (LRT) over all replicates for the 4 alternative analysis methods is presented in Figure 3 and Figure 4 for the 1H and AH phenotype simulations, for all 20 regions. In the analyses of the 1H phenotype, the results show that haplotype-based RGRM methods (HB1 and HB2) can capture the effect of an associated region better (i.e. produce a higher test statistic) than the SNP-based RGRM methods (SB1 and SB2). SNP-based RGRM methods can be used when the size of block is small (up to 7 SNPs) but, with increasing block size, the performance of these methods dropped. In the analyses of AH phenotypes, the results indicate that the haplotype-based RGRM methods perform better than SNP-based RGRM methods. Moreover, SNP-based RGRM methods fail to detect the effect of some regions (i.e. the test statistic is less than the Bonferroni correction for a significance level of 0.05) with increasing block size (i.e. over 12 SNPs). This is due to the fact that in the 1H scenario just one haplotype contains the associated effect and therefore SNP alleles just have the effect on traits when they are in the chosen haplotype and, in the AH scenario, all haplotypes have effects on the trait but each haplotype has a different effect regardless of the SNP alleles it carries. Therefore, SNP-based RGRM methods may not be successful in capturing the associated effect under these scenarios.

Use of haplotype blocks. Nagamine et al. (2012) used a SNP-based RGRM method where the analysis windows (regions) were defined using a fixed number of SNPs, ignoring the haplotype structure. Hence windows in that study may contain incomplete and/or several blocks. Therefore, in Nagamine et al. (2012) SNPs within a window that are used to construct the RGRM may be in low or high LD. In this study, we define windows on the basis of the local haplotype structure, and that allows us to better interpret the differences observed in the performance of the different analyses methods.

In the current study, we show that the haplotype-based RGRM method HB1 is more powerful than SNP-based RGRM methods to detect haplotype-based variants (as in 1H and AH), especially when the block size is larger.

CONCLUSIONS

The results of the current study suggest that haplotype-based methods can capture a proportion of the missing heritability explained by rare haplotypes. To detect SNP-based variants (as in 1S and AS), the SNP-based and the haplotype-based RGRM methods can be used. However, to detect haplotype-based variants haplotype-based RGRM methods should be used.

ACKNOWLEDGEMENTS

This work was supported by funding from the Biotechnology and Biological Sciences Research Council (UK) and the Medical Research Council (UK).

LITERATURE CITED

Aulchenko, Y.S., Ripke, S., Isaacs, A. et al. (2007). Bioinformatics. 23: 1294–1296.

Browning, S.R., and Browning, B.L. (2007). Am. J. Hum. Genet. 81:1084-1097.

Cebamanos, L., Gray, A., Stewart, I. et al. (2014). Bioinformatics. 30: btt754.

Nagamine, Y., Pong-Wong, R., Navarro, P. et al. (2012). PLOS ONE. 7: e46501.

Teslovich, T.M., Musunuru, K., Smith, A.V. et al. (2010). Nature 5; 466: 707-713.

VanRaden, P.M., (2007). INTERBULL Bull. 37: 111-114.

Yang, J., Benyamin, B., McEvoy, B.P. et al. (2010). Nat. Genet. 42: 565–569.